# Approximating and Reasoning about Data Provenance

**UNIVERSITY OF PENNSYLVANIA**

PI: IVES, ZACHARY (contact); KIM, JUNHYONG

Grant Number: 1 U01 EB020954-01

In many Big Data applications today, such as Next-Generation Sequencing, data processing pipelines are highly complex, span multiple institutions, and include many human and computational steps. The pipelines evolve over time and vary across institutions, so it is difficult to track and reason about the processing pipelines to ensure consistency and correctness of results. Provenance-enabled scientific workflow systems promise to aid here - yet such workflow systems are often avoided due to perceptions of inflexibility, lack of good provenance analytics tools, and emphasis on supporting the data consumer rather than producer. We propose to better incentivize the adoption of workflow and other provenance tracking tools: (1) Instead of requiring a single workflow system across the entire pipeline, which can be inflexible, we allow for integration across multiple autonomous systems (provenance- enabled workflow systems, provenance tracking systems for languages like Python and R, etc.), and even across steps performed without any provenance tracking at all. (2) We develop provenance reasoning capabilities specifically useful to the data provider, such as provenance analytics across time, sites, and users; finding the code modules that best explain why two results are different; regression testing to determine whether a code change would affect prior results; and reconstructing missing provenance for steps that were not captured. These capabilities are expected to lead to wider tracking of data provenance, and ultimately to more consistent, reproducible, and reliable science. We will validate this hypothesis through the evaluation of our technologies within a Next-Generation Sequencing pipeline run by one of the PIs with collaborators at other institutions. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: Settings like Next-Generation Sequencing have very complex data processing pipelines which change over time, making reasoning about data quality and consistency difficult. Data provenance tools promise to help in this respect, but are often viewed as burdensome and oriented towards the data consumer rather than producer. To incentivize adoption of provenance tracking and reasoning, we (1) make it lighter-weight to record and reconstruct the provenance of results in the data pipeline, (2) provide analytics and debugging tools over data provenance that help the data provider reconstruct missing provenance, understand changes, and troubleshoot unexpected differences in results.